From: Kristian G. Andersen
Sent: Mon, 27 Jul 2020 18:05:59 -0700
To: Fauci, Anthony (NIH/NIAID) [E]
Cc: Jeremy Farrar;Edward Holmes
Subject: Fwd: The authors who wrote the paper saying that SARS-CoV-2 is not human engineered first tried convincing Anthony Fauci of the opposite.
Attachments: Summary.Feb7.pdf

Dear Tony,

I am sorry to be contacting you, as I know you have critically important priorities, including developing a vaccine for COVID-19. **We just received the email below from Jon Cohen** (from Science) about our conversations back in February investigating the **origins of SARS-CoV-2**. As you know, we considered the theory that SARS-CoV-2 could have been a lab escape and therefore did what any good scientist should do - investigate likely hypotheses and let the data decide. As you know, the data strongly suggests that this is a natural virus and clearly this person gets a lot of things wrong about how this all played out.

We need to reply back to Jon, which would have to include confirming that this meeting did indeed take place with you and Jeremy present. Please let me know if you have any comments or concerns in this regard.

At the very end of this email, I have added a draft email that Eddie put together. I have a few clarifying points that I will add and then Eddie and I will reply back to Jon.

Again, sorry to take up your time - please let me know if you have any comments, questions, or concerns. **We are planning to email Jon tomorrow afternoon**.

Best,
Kristian

Kristian G. Andersen, PhD
Professor | Scripps Research
Director of Infectious Disease Genomics | Scripps Research Translational Institute
Vice President | Viral Hemorrhagic Fever Consortium
Principal Investigator | Center for Viral Systems Biology
Principal Investigator | West African Emerging Infectious Disease Research Center

The Scripps Research Institute
10550 North Torrey Pines Road, SGM-300A
Department of Immunology and Microbial Science
La Jolla, CA 92037

p: (858) 784-2118
t: @K_G_Andersen
e: andersen@scripps.edu
w: www.andersen-lab.com

Assistant: Michelle Platero, michelle@scripps.edu

---------- Forwarded message ---------
From: **Jon Cohen** <jcohen@aaas.org>
Date: Mon, Jul 27, 2020 at 3:02 PM
Subject: Re: The authors who wrote the paper saying that SARS-CoV-2 is not human engineered first tried convincing Anthony Fauci of the opposite.
To: Kristian G. Andersen [redacted] (b) (6), Edward Holmes <edward.holmes@sydney.edu.au>

Here's what one person who claims to have inside knowledge is saying behind your backs…
Jon

On Jul 25, 2020, at 7:22 AM, ofu8Iedu8z [redacted] (b) (6) wrote:

[EXTERNAL EMAIL]

Hello Jon
Given your recent mentions of the origin of SARS-CoV-2 I thought you might be interested to hear the bizarre back-story of the paper "The proximal origin of SARS-CoV-2" (https://www.nature.com/articles/s41591-020-0820-9).
In summary, four of the authors managed to organize a conference call with Anthony Fauci and others, after quietly raising the alarm (or "spreading the rumor", as Jeremy Farrar apparently put it) that the virus **WAS** in fact human engineered. On the call were two world-class virologists who actually work on coronaviruses, who set them straight in great detail. That seemed to be the end of the affair.
But, incredibly, Andersen et al. turned around and submitted the Proximal paper to *Nature* with the exact opposite claim, i.e., that the virus was **NOT** human engineered. They used (without acknowledgment, of course) all the arguments provided by the coronavirologists on the initial call in which they had tried to raise the human-engineered alarm.
I don't think it would be too hard to verify all this, if you feel like digging a little. If you're wondering if this could all possibly be true: ask yourself how this group of authors, none of whom work on coronaviruses, could have such detailed arguments about why SARS-CoV-2 was not human-engineered. The answer is that they couldn't (and didn't) - they were schooled by the coronavirus experts on the call.
For the phone conference, Anthony Fauci called in Jeremy Farrar (Director of the Wellcome Trust). Farrar asked the coronavirus experts to join the call to listen to the claims. The call took place on a Saturday in early February (either the 1st or 8th, I'm not sure but I could probably find out). On the call making the claim were: Kristian G. Andersen, Andrew Rambaut, Edward C. Holmes, Robert F. Garry, but not Ian Lipkin.
The coronavirus experts listened for a while and both quickly concluded that the reasoning was completely flawed, that the non-coronavirus virologists had no idea what they were talking about, and that the human-engineered claim was totally wrong. One of the coronavirus experts was entertaining guests that day and told the people on the conference call that they wanted to give their opinion and then go back to the guests. So they told them it was nonsense, gave them a list of reasons why, and got off the call. The other coronavirus expert stayed on the call, gave a

similar opinion and the morning afterwards sent a detailed list of the reasons why the claim was certainly wrong.

After the paper with the exact opposite claim was received at *Nature*, senior editor, Clare Thomas sent it out for review to some of the best people in the world... Not surprisingly, this happened to include a very close colleague of one of the experts who had been on the conference call. You can perhaps imagine the shock. Thomas was quickly appraised of the situation and *Nature* rejected the paper. It was then sent to *Nature Medicine*, where it was soon published.

One author on the paper was not on the conference call: Ian Lipkin. It's not clear how much of the back-story he is aware of. It might be worth giving him a call to ask, in case you feel like investigating. If his co-authors left him in the dark as to what actually happened and he's worried about the possible fallout he may want to help.

I apologize for mailing you without revealing my name (at least for now). I work in the field and have heard this story from two people who were on the initial call with Fauci. I'm not keen to be personally involved, but I find the situation so outrageous, hypocritical, and shameless that I also find I can't keep silent. It doesn't change anything with respect to knowledgeable thinking about the origin of the virus, of course, but it's a pretty ugly situation that I (obviously) think should be exposed.

----- EMAIL REPLY DRAFT -----

Hi Jon,

Here are the facts:

1. In early Feb we had spotted some features in the SARS-CoV-2 genome that at the time appeared unusual - particularly the furin cleavage site and the receptor binding domain.

2. At this stage we thought it was to wise ask for some other expert opinion on this, so a conference call was arranged. There were indeed some coronavirus experts on the call who we chose.

3. Clearly, some people on the call were very strongly of the opinion the possibility of a lab escape was ridiculous and listed reasons why it was unlikely (although there was also some initial confusion about whether we were referring to the crazy HIV origins theory that had just been touted - obviously we were not). Some of those comments we agreed with, others we didn't. There as a long email discussion about what the data said. A take-home message from the call was that we should go away and write something to clearly set-out the background science on the issue.

4. So, we eventually wrote up a paper. Critically, however, drafts of this paper were sent to all the people on the call, including those that have leaked out the information. I've attached here the draft of the document from Feb 7 that was circulated to everyone. As you can see, it is essentially the basis of the document and people on the call commented on it.

5. Very shortly after the call the pangolin data came out. This was critical. As I wrote in an email to everyone on the call on Feb 9th:

“Personally, with the pangolin virus possessing 6/6 key sites in the receptor binding domain, I am in favour of the natural evolution theory.”

6. Hence, it is completely and utterly false to claim that we all thought it was a lab escape, we were corrected in our views by the coronavirus experts on the call, and then submitted a Nature paper without anyone else knowing about it. The truth is that we had a range of views among us, our paper included the pangolin data that was not available at the time of the call, and we circulated drafts of our document to everyone.

I also strongly reject the idea that we should not have raised nor discussed the possibility of lab escape: as scientists we have to present all the data and discuss it openly. That’s all we did. To have not mentioned the possibility of lab escape would have been negligent. Is the person who emailed you seriously suggesting that we should have not discussed these issues? Wouldn’t that be a cover-up? Indeed, the great irony is that 99.9% of the feedback I’ve had on the paper - including death threats - are people accusing me of dismissing the lab escape theory too quickly!! Can you imagine if we had not mentioned it all?

This is clearly just case of sour grapes based on some half-truths. It’s telling that the person who emailed you is anonymous. I’ve absolutely no problem with people knowing that my views on this issue have evolved as more data have appeared. That’s science. Indeed, I’ve told this to many people: the way see it is that we set-up an hypothesis and then tested it. As far I can tell we are only ‘guilty’ of following the proper scientific method.

Hope this helps.

Eddie

Overview

Sequencing of 2019-nCoV revealed two notable features of its genome. We investigate these features and outline some examples for how the virus may have acquired them. We also discuss some scenarios by which these features could have arisen. **Analysis of the virus genome sequences clearly demonstrates that the virus is not a laboratory construct or experimentally manipulated virus**. We believe the features discussed, which may explain the infectiousness and transmissibility of 2019-nCoV in humans, could have arisen through selection and adaptation prior to the initial outbreak.

The two primary features of 2019-nCoV of interest were:

- Based on structural modeling and early biochemical experiments, 2019-nCoV appears to be optimized for binding to the human ACE2 receptor.

- The highly variable spike protein of 2019-nCoV has a furin cleavage inserted at the S1 and S2 boundary via the insertion of twelve in-frame nucleotides. Additionally, this event also led to the acquisition of three predicted O-linked glycans around the furin cleavage site.

Mutations in the receptor binding domain of 2019-nCoV

The receptor binding domain (RBD) in the spike protein of SARS-CoV and SARS-like coronaviruses is the most variable part of the virus genome. When aligned against related viruses, 2019-nCoV displays a similar level of diversity as predicted from previous studies, including to its most closely related virus - SARS-like CoV isolated from bats (RaTG13, which is ~96% identical to 2019-nCoV).

Six residues in the RBD have been described as critical for binding to the human ACE2 receptor and determining host range[1]. Using coordinates based on the Ubani strain of SARS-CoV, they are Y442, L472, N479, D480, T487, and Y491 (the corresponding residues in 2019-nCoV are L455, F486, Q493, S494, N501, and Y505). Five out of six of these residues are mutated in 2019-nCoV compared to the closely related virus, RaTG13 (**Figure 1**). Based on modeling[1] and early biochemical experiments[2,3], 2019-nCoV seems to have an RBD that may bind with high affinity to ACE2 from human, primate, ferret, pig, and cat, as well as other species with high receptor homology. In contrast, 2019-nCoV may bind less efficiently to ACE2 in other species associated with SARS-like viruses, including rodents, civets, and bats[1].

A phenylalanine at F486 in 2019-nCoV corresponds to L472 in the SARS-CoV Ubani strain. In cell culture experiments the leucine at position 472 mutated to phenylalanine (L472F)[4], which has been predicted to be optimal for binding of the SARS-CoV RBD to the human ACE2 receptor[5]. However, a phenylalanine in this position is also present in several SARS-like CoVs from bats (**Figure 1**). While these analyses suggest that 2019-nCoV may be capable of binding the human ACE2 receptor with high affinity, importantly, the interaction is not predicted to be optimal[1]. Additionally, several of the key residues in the RBD of 2019-nCoV are different from those previously described to be optimal for human ACE2 receptor binding as determined by both natural evolution of SARS-CoV and rational design[5]. This latter point is strong evidence *against* 2019-nCoV being specifically engineered as, presumably, in such a scenario the most optimal residues would have been introduced, which is not what we observe.

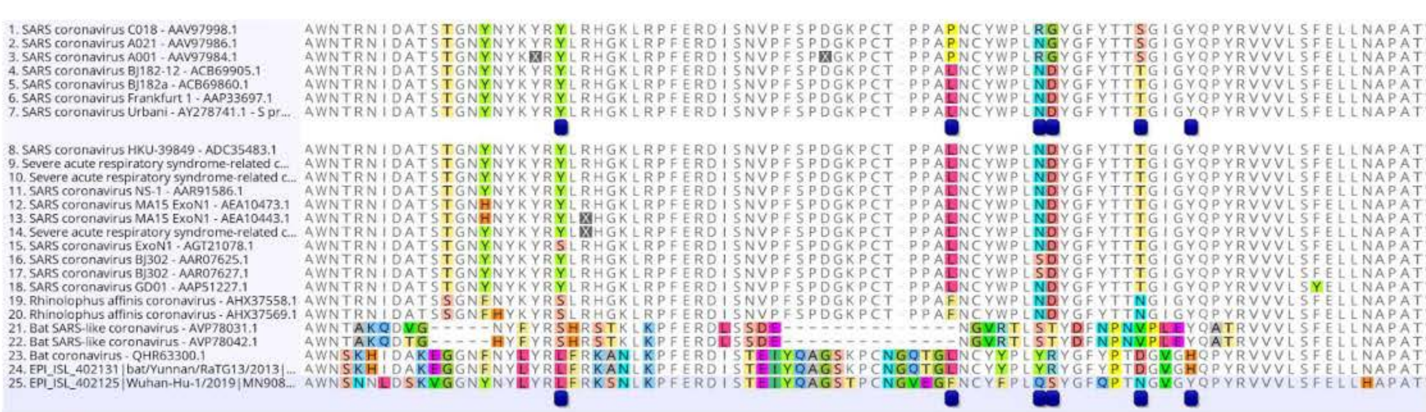


Figure 1 | Mutations in contact residues of the 2019-nCoV spike protein. The spike protein of 2019-nCoV (bottom) was aligned against the most closely related SARS and SARS-like CoVs. Key residues in the spike protein that make contact to the

ACE2 receptor have been marked with blue boxes in both 2019-nCoV and the SARS-CoV Urbani strain.

Furin cleavage site and O-linked glycans

An interesting feature of 2019-nCoV is a predicted furin cleavage site in the spike protein (**Figure 2**). In addition to the furin cleavage site (RRAR), a leading P is also inserted so the fully inserted sequence becomes PRRA (**Figure 2**). A proline in this position is predicted to create three flanking O-linked glycans at S673, T678, and S686. A furin site has never before been observed in the lineage B betacoronaviruses and is a unique feature of 2019-nCoV. Some human betacoronaviruses, including HCoV-HKU1 (lineage A) have furin cleavage sites (typically RRKR), although not in such an optimal position.

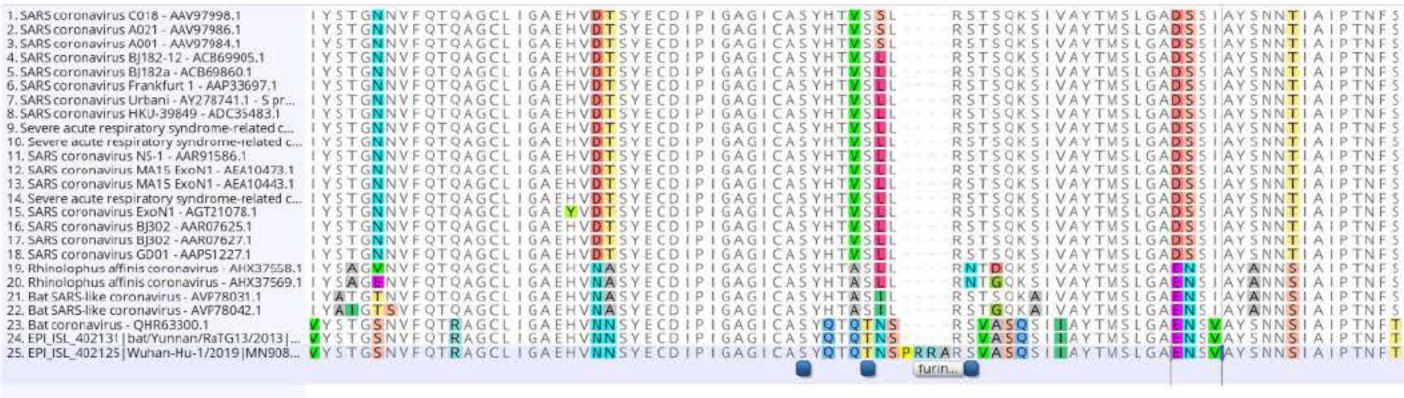


Figure 2 | Acquisition of furin cleavage site and O-linked glycans. The spike protein of 2019-nCoV (bottom) was aligned against the most closely related SARS and SARS-like CoVs. The furin cleavage site is marked in grey with the three adjacent predicted O-linked glycans in blue. Both the furin cleavage site and O-linked glycans are unique to 2019-nCoV and not previously seen in this group of viruses.

While the functional consequence - if any - of the furin cleavage site in 2019-nCoV is unknown, previous experiments with SARS-CoV have shown that it enhances cell–cell fusion but does not affect virus entry[6]. Furin cleavage sites are often acquired in condition selecting for rapid virus replication and transmission (e.g., highly dense chicken populations) and are a hallmark of highly pathogenic avian influenza virus, although these viruses acquire the site in different and more direct ways[7–9]. The acquisition of furin cleavage sites have also been observed after repeated passage of viruses in cell culture (personal correspondence and NASEM call, February 3, 2020).

A potential function of the three predicted O-linked glycans is less clear, but could create a “mucin-like domain” shielding potential epitopes or key residues on the 2019-nCoV spike protein.

Origin of 2019-nCoV

As noted at the start of this document, we believe that the origin of 2019-nCoV through laboratory manipulation of an existing SARS-related coronavirus can be ruled out with a high degree of confidence. If genetic manipulation would have been performed, one would expect that a researcher would have used one of the several reverse genetics systems available for betacoronaviruses. However, this is not the case as the genetic data clearly shows that 2019-nCoV is not derived from any previously used virus backbone, for example those described in a 2015 paper in *Nature Medicine*[10].

Instead we believe one of three main scenarios could explain how 2019-nCoV acquired the features discussed above: (1) natural selection in humans, (2) natural selection in an animal host, or (3) selection during passage.

Adaptation to humans

As the features outlined above are likely to enhance the ability of the virus to infect humans, it is possible that these are indeed adaptations to humans as a host and arose after the virus jumped from a non-human host, during the early stages of the epidemic. However, all of the genome sequences so far have the features described above and estimates of the timing of the most recent common ancestor of the currently sampled viruses support the seafood market outbreak as the zoonotic origin (i.e., in early December) and this would afford little opportunity for adaptation to occur. This may be explained by a transition to a rapid growth phase in the epidemic when the features arose and from which all current

cases are derived. However this would require a prior hidden epidemic of sufficient magnitude and duration for the adaptations to occur and there is no evidence of this. We also note that these features did not emerge during the SARS epidemic, which involved extensive human to human transmission.

Selection in an animal host

Given the similarity of 2019-nCoV to bat SARS-like CoVs, particularly RaTG13, it is highly likely that bats serve as the reservoir for this virus. However, previous human epidemics caused by betacoronaviruses have involved intermediate (possibly amplifying) hosts such as civets and other animals (SARS) and camels (MERS). It is therefore likely that an intermediate host would also exist for 2019-nCoV, although it is unclear what that host may be. Given the mutations in key residues of the RBD in 2019-nCoV it seems less likely that civets would be involved, although it is impossible to say with certainty at this stage. Notably, provisional analyses reveal that Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain CoVs that are extremely similar to 2019-nCoV[11]. Although RaTG13 remains the closest relative to 2019-nCoV across the genome as a whole, the Malayan pangolin CoVs are identical to 2019-nCoV at all six key RBD residues. Analyses of these pangolin viruses are ongoing, although they do not carry the furin cleavage site insertion.

For the virus to acquire the furin cleavage site and mutations in the spike proteins that appear to be suitable for human ACE2 receptor binding, it seems plausible that this animal host would have to have a high population density – to allow the necessary natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue. Since furin cleavage sites have not been observed in sarbecoviruses before, it is unclear what conditions would be required for it to be acquired in the lineage leading to 2019-nCoV.

Selection during passage

Basic research involving passage of bat SARS-like coronaviruses in cell culture and/or animal models have been ongoing in BSL-2 for many years across the world, including in Wuhan (e.g.,[12–15]). It is possible that 2019-nCoV could have acquired the RBD mutations and furin cleavage site as part of passage in cell culture, which have been observed in previous studies with e.g., SARS-CoV[4]. However, it is less clear how the O-linked glycans - if functional - would have been acquired, as these typically suggest the involvement of an immune system, which is not present *in vitro*. In this scenario, it is also unclear how the virus would be linked to the fact that the epidemic seemed to 'take off' at a particular food market, although the exact role of this locality is currently uncertain.

Limitations and recommendations

The evolution scenarios discussed above are largely indistinguishable and current data are consistent with all three. It is currently impossible to prove or disprove either, and it is unclear whether future data or analyses will help resolve this issue. Identifying the immediate non-human animal source and obtaining virus sequences from it would be the most definitive way of distinguishing the three scenarios.

The main limitation of what is described here is our clear ascertainment bias. We are looking for features or evolutionary aspects that could help explain how 2019-nCoV lead to such a rapidly expanding human epidemic, yet the specific features we are trying to find may be the exact features one would expect in a virus that could lead to an epidemic of the magnitude currently observed. Before 2019-nCoV 'took off' and started the current epidemic, it is plausible that many stuttering transmission chains of highly similar viruses could have entered the human population, but because they never took off they were never sampled. It is extremely important to keep this in mind as any inference about the plausibility of various scenarios about the evolution and/or epidemic potential of 2019-nCoV is attempted.

To further clarify the evolutionary origins and functional features of 2019-nCoV it would be helpful to obtain additional data about the virus - both genetic and functional. This includes experimental studies of receptor binding and the role of the furin cleavage site and predicted O-linked glycans. The identification of a potential intermediate host of 2019-nCoV as well as sequencing of very early cases, including those not connected to the market, could also help refute the passage scenario described above. Even in the

light of such data, however, it is not guaranteed that data can be obtained to conclusively prove all aspects of the initial emergence of 2019-nCoV.

References

1. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J. Virol.* (2020) doi:10.1128/JVI.00127-20.
2. Letko, M. & Munster, V. Functional assessment of cell entry and receptor usage for lineage B β-coronaviruses, including 2019-nCoV. *bioRxiv* 2020.01.22.915660 (2020) doi:10.1101/2020.01.22.915660.
3. Hoffmann, M. *et al.* The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *bioRxiv* 2020.01.31.929042 (2020) doi:10.1101/2020.01.31.929042.
4. Sheahan, T. *et al.* Mechanisms of zoonotic severe acute respiratory syndrome coronavirus host range expansion in human airway epithelium. *J. Virol.* **82**, 2274–2285 (2008).
5. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
6. Follis, K. E., York, J. & Nunberg, J. H. Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology* **350**, 358–369 (2006).
7. Longping, V. T., Hamilton, A. M., Friling, T. & Whittaker, G. R. A novel activation mechanism of avian influenza virus H9N2 by furin. *J. Virol.* **88**, 1673–1683 (2014).
8. Alexander, D. J. & Brown, I. H. History of highly pathogenic avian influenza. *Rev. Sci. Tech.* **28**, 19–38 (2009).
9. Luczo, J. M. *et al.* Evolution of high pathogenicity of H5 avian influenza virus: haemagglutinin cleavage site selection of reverse-genetics mutants during passage in chickens. *Sci. Rep.* **8**, 11518 (2018).
10. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).
11. virological.org: http://virological.org/t/ncov-2019-spike-protein-receptor-binding-domain-shares-high-amino-acid-identity-with-a-coronavirus-recovered-from-a-pangolin-viral-metagenomic-dataset/362 (2020).
12. Ge, X.-Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535–538 (2013).
13. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**, e1006698 (2017).
14. Zeng, L.-P. *et al.* Bat Severe Acute Respiratory Syndrome-Like Coronavirus WIV1 Encodes an Extra Accessory Protein, ORFX, Involved in Modulation of the Host Immune Response. *J. Virol.* **90**, 6573–6582 (2016).

15. Yang, X.-L. *et al.* Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. *J. Virol.* **90**, 3253–3256 (2015).

From: Fauci, Anthony (NIH/NIAID) [E]
Sent: Sun, 8 Mar 2020 13:23:32 +0000
To: Kristian G. Andersen
Cc: Jeremy Farrar;Collins, Francis (NIH/OD) [E];Robert Garry;Edward Holmes;Andrew Rambaut;Ian Lipkin;Chris Emery
Subject: Re: SARS-CoV-2 article to be published in Nature Medicine

Kristian:
Thanks for your note. Nice job on the paper.
Tony

On Mar 6, 2020, at 4:23 PM, Kristian G. Andersen <andersen@scripps.edu> wrote:

Dear Jeremy, Tony, and Francis,

Thank you again for your advice and leadership as we have been working through the SARS-CoV-2 'origins' paper. We're happy to say that the paper was just accepted by Nature Medicine and should be published shortly (not quite sure when).

To keep you in the loop, I just wanted to share the accepted version with you, as well as a draft press release. We're still waiting for proofs, so please let me know if you have any comments, suggestions, or questions about the paper or the press release.

Tony, thank you for your straight talk on CNN last night - it's being noticed.

Best,
Kristian

Kristian G. Andersen, PhD
Associate Professor, Scripps Research
Director of Infectious Disease Genomics, Scripps Research Translational Institute
Director, Center for Viral Systems Biology

The Scripps Research Institute
10550 North Torrey Pines Road, SGM-300A
Department of Immunology and Microbial Science
La Jolla, CA 92037

p: (858) 784-2118
c: [redacted] (b) (6)
t: @K_G_Andersen
e: andersen@scripps.edu
w: www.andersen-lab.com

Assistant: Michelle Platero, michelle@scripps.edu

<Andersen Coronavirus Nature 2020 Press Release Draft 4.docx>
<Manuscript.pdf>

From: Kristian G. Andersen
Sent: Fri, 6 Mar 2020 13:22:28 -0800
To: Jeremy Farrar;Fauci, Anthony (NIH/NIAID) [E];Collins, Francis (NIH/OD) [E]
Cc: Robert Garry;Edward Holmes;Andrew Rambaut;Ian Lipkin;Chris Emery
Subject: SARS-CoV-2 article to be published in Nature Medicine
Attachments: Andersen Coronavirus Nature 2020 Press Release Draft 4.docx, Manuscript.pdf

Dear Jeremy, Tony, and Francis,

Thank you again for your advice and leadership as we have been working through the SARS-CoV-2 'origins' paper. We're happy to say that the paper was just accepted by Nature Medicine and should be published shortly (not quite sure when).

To keep you in the loop, I just wanted to share the accepted version with you, as well as a draft press release. We're still waiting for proofs, so please let me know if you have any comments, suggestions, or questions about the paper or the press release.

Tony, thank you for your straight talk on CNN last night - it's being noticed.

Best,
Kristian

Kristian G. Andersen, PhD
Associate Professor, Scripps Research
Director of Infectious Disease Genomics, Scripps Research Translational Institute
Director, Center for Viral Systems Biology

The Scripps Research Institute
10550 North Torrey Pines Road, SGM-300A
Department of Immunology and Microbial Science
La Jolla, CA 92037

p: (858) 784-2118
c: (b) (6)
t: @K_G_Andersen
e: andersen@scripps.edu
w: www.andersen-lab.com

Assistant: Michelle Platero, michelle@scripps.edu

Andersen Coronavirus Nature Medicine Press Release Draft 2-24-20

The COVID-19 coronavirus epidemic has a natural origin, scientists say

The novel SARS-CoV-2 coronavirus that emerged in the city of Wuhan, China, last year and has since caused a large scale COVID-19 epidemic and spread to more than 70 other countries is the product of natural evolution, according to findings published today in the journal *Nature Medicine*.

The analysis of public genome sequence data from SARS-CoV-2 and related viruses found no evidence that the virus was made in a laboratory or otherwise engineered.

"By comparing the available genome sequence data for known coronavirus strains, we can firmly determine that SARS-CoV-2 originated through natural processes," said Kristian Andersen, PhD, an associate professor of immunology and microbiology at Scripps Research and corresponding author on the paper.

In addition to Andersen, authors on the paper include Robert F. Garry, of Tulane University; Edward Holmes, of the University of Sydney; Andrew Rambaut, of University of Edinburgh; W. Ian Lipkin, of Columbia University.

Coronaviruses are a large family of viruses that can cause illnesses ranging widely in severity. The first known severe illness caused by a coronavirus emerged with the 2003 Severe Acute Respiratory Syndrome (SARS) epidemic in China. A second outbreak of severe illness began in 2012 in Saudi Arabia with the Middle East Respiratory Syndrome (MERS).

On December 31 of last year, Chinese authorities alerted the World Health Organization of an outbreak of a novel strain of coronavirus causing severe illness, which was subsequently named SARS-CoV-2. As of February 20, 2020, nearly 100,000[TBD] COVID-19 cases have been documented, although many more mild cases have likely gone undiagnosed. The virus has killed over 3,000[TBD] people.

Shortly after the epidemic began, Chinese scientists sequenced the genome of SARS-CoV-2 and made the data available to researchers worldwide. The resulting genomic sequence data has shown that Chinese authorities rapidly detected the epidemic and that the number of COVID-19 cases have been increasing because of human to human transmission after a single introduction into the human population. Andersen and collaborators at several other research institutions used this sequencing data to explore the origins and evolution of SARS-CoV-2 by focusing in on several tell-tale features of the virus.

The scientists analyzed the genetic template for spike proteins, armatures on the outside of the virus that it uses to grab and penetrate the outer walls of human and animal cells. More specifically, they focused on two important features of the spike protein: the receptor-binding

domain (RBD), a kind of grappling hook that grips onto host cells, and the cleavage site, a molecular can opener that allows the virus to crack open and enter host cells.

Evidence for natural evolution

The scientists found that the RBD portion of the SARS-CoV-2 spike proteins had evolved to effectively target a molecular feature on the outside of human cells called ACE2, a receptor involved in regulating blood pressure. The SARS-CoV-2 spike protein was so effective at binding the human cells, in fact, that the scientists concluded it was the result of natural selection and not the product of genetic engineering.

This evidence for natural evolution was supported by data on SARS-CoV-2's backbone – its overall molecular structure. If someone were seeking to engineer a new coronavirus as a pathogen, they would have constructed it from the backbone of a virus known to cause illness. But the scientists found that the SARS-CoV-2 backbone differed substantially from those of already known coronaviruses and mostly resembled related viruses found in bats and pangolins.

"These two features of the virus, the mutations in the RBD portion of the spike protein and its distinct backbone, rules out laboratory manipulation as a potential origin for SARS-CoV-2" said Andersen.

Josie Golding, PhD, epidemics lead at UK-based Wellcome Trust, said the findings by Andersen and his colleagues are "crucially important to bring an evidence-based view to the rumors that have been circulating about the origins of the virus (SARS-CoV-2) causing COVID-19."

"They conclude that the virus is the product of natural evolution," Goulding adds, "ending any speculation about deliberate genetic engineering."

Possible origins of the virus

Based on their genomic sequencing analysis, Andersen and his collaborators concluded that the most likely origins for SARS-CoV-2 followed one of two possible scenarios.

In one scenario, the virus evolved to its current pathogenic state through natural selection in a non-human host and then jumped to humans. This is how previous coronavirus outbreaks have emerged, with humans contracting the virus after direct exposure to civets (SARS) and camels (MERS). The researchers proposed bats as the most likely reservoir for SARS-CoV-2 as it is very similar to a bat coronavirus. There are no documented cases of direct bat-human transmission, however, suggesting that an intermediate host was likely involved between bats and humans.

In this scenario, both of the distinctive features of SARS-CoV-2's spike protein—the RBD portion that binds to cells and the cleavage site that opens the virus up—would have evolved to their current state prior to entering humans. In this case, the current epidemic would probably have emerged rapidly as soon as humans were infected, as the virus would have already evolved the features that make it pathogenic and able to spread between people.

In the other proposed scenario, a non-pathogenic version of the virus jumped from an animal host into humans and then evolved to its current pathogenic state within the human population. For instance, some coronaviruses from pangolins, armadillo-like mammals found in Asia and Africa, have an RBD structure very similar to that of SARS-CoV-2. A coronavirus from a pangolin could possibly have been transmitted to a human, either directly or through an intermediary host such as civets or ferrets.

Then the other distinct spike protein characteristic of SARS-CoV-2, the cleavage site, could have evolved within a human host, possibly via limited undetected circulation in the human population prior to the beginning of the epidemic. The researchers found that the SARS-CoV-2 cleavage site, appears similar to the cleavage sites of strains of bird flu that has been shown to transmit easily between people. SARS-CoV-2 could have evolved such a virulent cleavage site in human cells and soon kicked off the current epidemic, as the coronavirus would possibly have become far more capable of spreading between people.

Study co-author Andrew Rambaut cautioned that it is difficult if not impossible to know at this point which of the scenarios is most likely. If the SARS-CoV-2 entered humans in its current pathogenic form from an animal source, it raises the probability of future outbreaks, as the illness-causing strain of the virus could still be circulating in the animal population and might once again jump into humans. The chances are lower of a non-pathogenic coronavirus entering the human population and then evolving properties similar to SARS-CoV-2.

Funding for the research was provided by the US National Institutes of Health, the Pew Charitable Trusts, the Wellcome Trust, the European Research Council, and an ARC Australian Laureate Fellowship.

The Proximal Origin of SARS-CoV-2

Kristian G. Andersen[1,2*], Andrew Rambaut[3], W. Ian Lipkin[4], Edward C. Holmes[5] & Robert F. Garry[6,7]

[1]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA 92037, USA.
[2]Scripps Research Translational Institute, La Jolla, CA 92037, USA.
[3]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK.
[4]Center for Infection and Immunity, Mailman School of Public Health of Columbia University, New York, New York, USA.
[5]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia.
[6]Tulane University, School of Medicine, Department of Microbiology and Immunology, New Orleans, LA, USA.
[7]Zalgen Labs, LCC, Germantown, MD, USA.

*Corresponding author: andersen@scripps.edu

TO THE EDITOR - Since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China[1,2] there has been considerable discussion on the origin of the causative virus SARS-CoV-2[3] (also referred to as HCoV-19)[4]. Infections with SARS-CoV-2 are now widespread, and as of 29 February 2020, 86,012 cases have been confirmed in more than 60 countries, with 2,941 deaths[5].

SARS-CoV-2 is the seventh coronavirus known to infect humans. SARS-CoV, MERS-CoV, and SARS-CoV-2 can cause severe disease, whereas HKU1, NL63, OC43 and 229E, are associated with mild symptoms[6]. Herein, we review what can be deduced about the origin of SARS-CoV-2 from the comparative analysis of genomic data. We offer a perspective on the notable features in the SARS-CoV-2 genome and discuss scenarios by which they could have arisen. Our analyses clearly show that SARS-CoV-2 is not a laboratory construct nor a purposefully manipulated virus.

Notable features of the SARS-CoV-2 genome

Our comparison of alpha- and betacoronaviruses identifies two notable genomic features of SARS-CoV-2: (***i***) based on structural studies[7–9] and biochemical experiments[1,9,10], SARS-CoV-2 appears optimized for binding to the human ACE2 receptor; (***ii***) the spike (S) protein of SARS-CoV-2 has a functional polybasic (furin) cleavage site at the S1/S2 boundary through the insertion of twelve nucleotides[8]. Additionally, this led to the predicted acquisition of three O-linked glycans around the site.

1. Mutations in the receptor binding domain of SARS-CoV-2

The receptor binding domain (RBD) in the spike protein is the most variable part of the coronavirus genome[1,2]. Six RBD amino acids have been shown to be critical for binding to ACE2 receptors and determining the host range of SARS-like viruses[7]. Using coordinates based on SARS-CoV, they are Y442, L472, N479, D480, T487, and Y4911 corresponding to L455, F486, Q493, S494, N501, and Y505 in SARS-CoV-2[7]. Five of these six residues differ between SARS-CoV-2 and SARS-CoV (**Fig. 1a**). Based on structural studies[7–9] and biochemical experiments[1,9,10], SARS-CoV-2 seems to have an RBD that binds with high affinity to ACE2 from human, ferret, cat, and other species with high receptor homology[7].

While these analyses suggest that SARS-CoV-2 may bind human ACE2 with high affinity, computational analyses predict that the interaction is not ideal[7] and the RBD sequence is different from those shown in SARS-CoV to be optimal for receptor binding[7,11]. Thus, the high affinity binding of the SARS-CoV-2 spike protein to human ACE2 is most likely the result of natural selection on a human or human-like ACE2 permitting another optimal binding solution to arise. This is strong evidence that SARS-CoV-2 is *not* the product of purposeful manipulation.

2. Polybasic furin cleavage site and O-linked glycans

The second notable feature of SARS-CoV-2 is a polybasic cleavage site (RRAR) at the S1/S2 junction, the two subunits of the spike (**Fig. 1b**)[8]. This allows effective cleavage by furin and other proteases and plays a role in determining virus infectivity and host range[12]. In addition, a leading proline is also inserted at this site in SARS-CoV-2; thus, the inserted sequence is PRRA (**Fig. 1b**). The turn created by the proline is predicted to result in the addition of O-linked glycans to S673, T678, and S686 flanking the cleavage site and are unique to SARS-CoV-2 (**Fig. 1b**). Polybasic cleavage sites have not been observed in related

"lineage B" betacoronaviruses, although other human betacoronaviruses, including HKU1 (lineage A), have them and predicted O-linked glycans[13]. Given the level of genetic variation in the spike it is likely that SARS-CoV-2-like viruses with partial or full polybasic cleavage sites will be discovered in other species.

The functional consequence of the polybasic cleavage site in SARS-CoV-2 is unknown and it will be important to determine its impact on transmissibility and pathogenesis in animal models. Experiments with SARS-CoV have shown that insertion of a furin cleavage site at the S1/S2 junction enhances cell–cell fusion without affecting virus entry[14]. In addition, efficient cleavage of the MERS-CoV spike enables MERS-like coronaviruses from bats to infect human cells[15]. In avian influenza viruses, rapid replication and transmission in highly dense chicken populations selects for the acquisition of polybasic cleavage sites in the haemagglutinin (HA) protein[16], which serves a similar function as the coronavirus spike protein. Acquisition of polybasic cleavage sites in HA, by insertion or recombination, converts low pathogenicity avian influenza viruses into highly pathogenic forms[16]. The acquisition of polybasic cleavage sites by HA has also been observed after repeated passage in cell culture or through animals[17].

The function of the predicted O-linked glycans is unclear, but they could create a "mucin-like domain" shielding epitopes or key residues on the SARS-CoV-2 spike protein[18]. Several viruses employ mucin-like domains as glycan shields involved in immune evasion[18]. Although prediction of O-linked glycosylation is robust, experimental studies are required to determine if these sites are utilized in SARS-CoV-2.

Theories of SARS-CoV-2 origins

It is improbable that SARS-CoV-2 emerged through laboratory manipulation of a related SARS-like coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for human ACE2 binding with an efficient solution different from those previously predicted[7,11]. Further, had genetic manipulation had been performed, one of the several reverse genetic systems available for betacoronaviruses would likely have been used[19]. However, the genetic data irrefutably show that SARS-CoV-2 is not derived from any previously used virus backbone[20]. Instead, we propose two scenarios that can plausibly explain the origin of SARS-CoV-2: (***i***) natural selection in an animal host prior to zoonotic transfer, and (***ii***) natural selection in humans following zoonotic transfer. We also discuss whether selection during passage could have given rise to SARS-CoV-2.

1. Natural selection in an animal host prior to zoonotic transfer

As many early cases of COVID-19 were linked to the Huanan market in Wuhan[1,2], it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-like coronaviruses[2], it is likely that bats serve as reservoir hosts for its progenitor. Although RaTG13, sampled from a *Rhinolophus affinis* bat[1], is ~96% identical overall to SARS-CoV-2, its spike diverges in the RBD suggesting that it may not bind efficiently to the human ACE2 receptor (**Fig. 1a**)[7].

Malayan pangolins (*Manis javanica*) illegally imported into Guangdong province contain coronaviruses similar to SARS-CoV-2[21]. Although the RaTG13 bat virus remains the closest relative to SARS-CoV-2 across the genome[1], some pangolin coronaviruses exhibit strong similarity to SARS-CoV-2 in the RBD, including all six key RBD residues (**Fig. 1**)[21]. This clearly shows that the SARS-CoV-2 spike protein optimized for binding to human-like ACE2 is the result of natural selection.

Neither the bat nor pangolin betacoronaviruses sampled to date have polybasic cleavage sites. Although no animal coronavirus has been identified that is sufficiently similar to have served as the direct SARS-CoV-2 progenitor, the diversity of coronaviruses in bats and other species is massively undersampled. Mutations, insertions and deletions, can occur near the S1/S2 junction of coronaviruses[22] showing that the polybasic cleavage site can arise by a natural evolutionary process. For a precursor virus to acquire both the polybasic cleavage site and mutations in the spike protein suitable for human ACE2 receptor binding, an animal host would likely have to have a high population density – to allow natural selection to proceed efficiently – and an ACE2 gene that is similar to the human orthologue.

2. Natural selection in humans following zoonotic transfer

It is possible that a progenitor to SARS-CoV-2 jumped into humans, acquiring the genomic features described above through adaptation during undetected human-to-human transmission. Once acquired,

these adaptations would enable the epidemic to take off, producing a sufficiently large cluster of cases to trigger the surveillance system that detected it[1,2].

All SARS-CoV-2 genomes sequenced so far have the genomic features derived above and are thus derived from a common ancestor that had them too. The presence in pangolins of an RBD very similar to that in SARS-CoV-2 means we can infer this was also likely in the virus that jumped to humans. This leaves the polybasic cleavage site insertion to occur during human-to-human transmission.

Estimates of the timing of the most recent common ancestor of SARS-CoV-2 using current sequence data point to virus emergence in late November to early December 2019[23], compatible with the earliest retrospectively confirmed cases[24]. Hence, this scenario presumes a period of unrecognised transmission in humans between the initial zoonotic event and the acquisition of the polybasic cleavage site. Sufficient opportunity could occur if there had been many prior zoonotic events producing short chains of human-to-human transmission over an extended period. This is essentially the situation for MERS-CoV where all human cases are the result of repeated jumps of the virus from dromedary camels, producing single infections or short transmission chains that eventually resolve, with no adaptation to sustained transmission[25].

Studies of banked human samples could provide information on whether such cryptic spread has occurred. Retrospective serological studies could also be informative and a few such studies have been conducted showing low-level exposures to SARS-like coronaviruses in certain areas of China [26]. Critically, however, these studies could not have distinguished whether exposures were due to prior infections with SARS-CoV, SARS-CoV-2, or other SARS-like coronaviruses. Further serological studies should be conducted to determine the extent of prior human exposure to SARS-CoV-2.

3. Selection during passage

Basic research involving passage of bat SARS-like coronaviruses in cell culture and/or animal models have been ongoing in BSL-2 for many years in laboratories across the world[27] and there are documented instances of laboratory escapes of SARS-CoV[28]. We must therefore examine the possibility of a inadvertent laboratory release of SARS-CoV-2.

In theory, it is possible that SARS-CoV-2 acquired RBD mutations (**Fig. 1a**) during adaptation to passage in cell culture, as has been observed in studies with SARS-CoV[11]. The finding of SARS-like coronaviruses from pangolins with near-identical RBDs, however, provides a much stronger and parsimonious explanation for how SARS-CoV-2 acquired these via recombination or mutation[19].

The acquisition of both the polybasic cleavage site and predicted O-linked glycans also argues against culture-based scenarios. New polybasic cleavage sites have only been observed after prolonged passage of low pathogenicity avian influenza virus *in vitro* or *in vivo*[17]. Furthermore, a hypothetical generation of SARS-CoV-2 by cell culture or animal passage would have required prior isolation of a progenitor virus with very high genetic similarity, which has not been described. Subsequent generation of a polybasic cleavage site would have then required repeated passage in cell culture or animals with ACE2 receptors similar to humans, but such work has also not previously been described. Finally, the generation of the predicted O-linked glycans is also unlikely to have occured due to cell culture passage, as such features suggest the involvement of an immune system[18].

Conclusions

In the midst of the global COVID-19 public health emergency it is reasonable to wonder why the origins of the epidemic matter. A detailed understanding of how an animal virus jumped species boundaries to infect humans so productively will help in the prevention of future zoonotic events. For example, if SARS-CoV-2 pre-adapted in another animal species then we are at risk of future re-emergence events. In contrast, if the adaptive process occurred in humans, then even if we have repeated zoonotic transfers they are unlikely to take off without the same series of mutations. In addition, identifying the closest animal relatives of SARS-CoV-2 will greatly assist studies of virus function. Indeed, the availability of the RaTG13 bat sequence helped reveal key RBD mutations and the polybasic cleavage site.

The genomic features described here may in part explain the infectiousness and transmissibility of SARS-CoV-2 in humans. Although the evidence shows that SARS-CoV-2 is not a purposefully manipulated virus, it is currently impossible to prove or disprove the other theories of its origin described here. However, since we observe all notable SARS-CoV-2 features - including the optimized RBD and polybasic cleavage site - in related coronaviruses in nature, we do not believe that any type of laboratory-based scenario is plausible.

More scientific data could swing the balance of evidence to favor one hypothesis over another. Obtaining related virus sequences from animal sources would be the most definitive way of revealing virus origins. For example, a future observation of an intermediate or fully formed polybasic cleavage site in an SARS-CoV-2-like virus from animals would lend even further support to the natural selection hypotheses. It would also be helpful to obtain more genetic and functional data about SARS-CoV-2, including animal studies. The identification of a potential intermediate host of SARS-CoV-2, as well as the sequencing of very early cases would similarly be highly informative. Irrespective of the exact mechanisms of how SARS-CoV-2 originated via natural selection, the ongoing surveillance of pneumonia in humans and other animals is clearly of utmost importance.

Acknowledgements

We thank all those who have contributed sequences to the GISAID database (https://www.gisaid.org/) and analyses to Virological.org (http://virological.org/). We thank M. Farzan for discussions. We thank the Wellcome Trust for support. KGA is a Pew Biomedical Scholar and is supported by NIH grant U19AI135995. AR is supported by the Wellcome Trust (Collaborators Award 206298/Z/17/Z – ARTIC network) and the European Research Council (grant agreement no. 725422 – ReservoirDOCS). ECH is supported by an ARC Australian Laureate Fellowship (FL170100022). RFG is supported by NIH grants U19AI135995, U54 HG007480 and U19AI142790.

Competing Interests

RFG is co-founder of Zalgen Labs, a biotechnology company developing countermeasures to emerging viruses. None of the other authors declare any conflicts of interest.

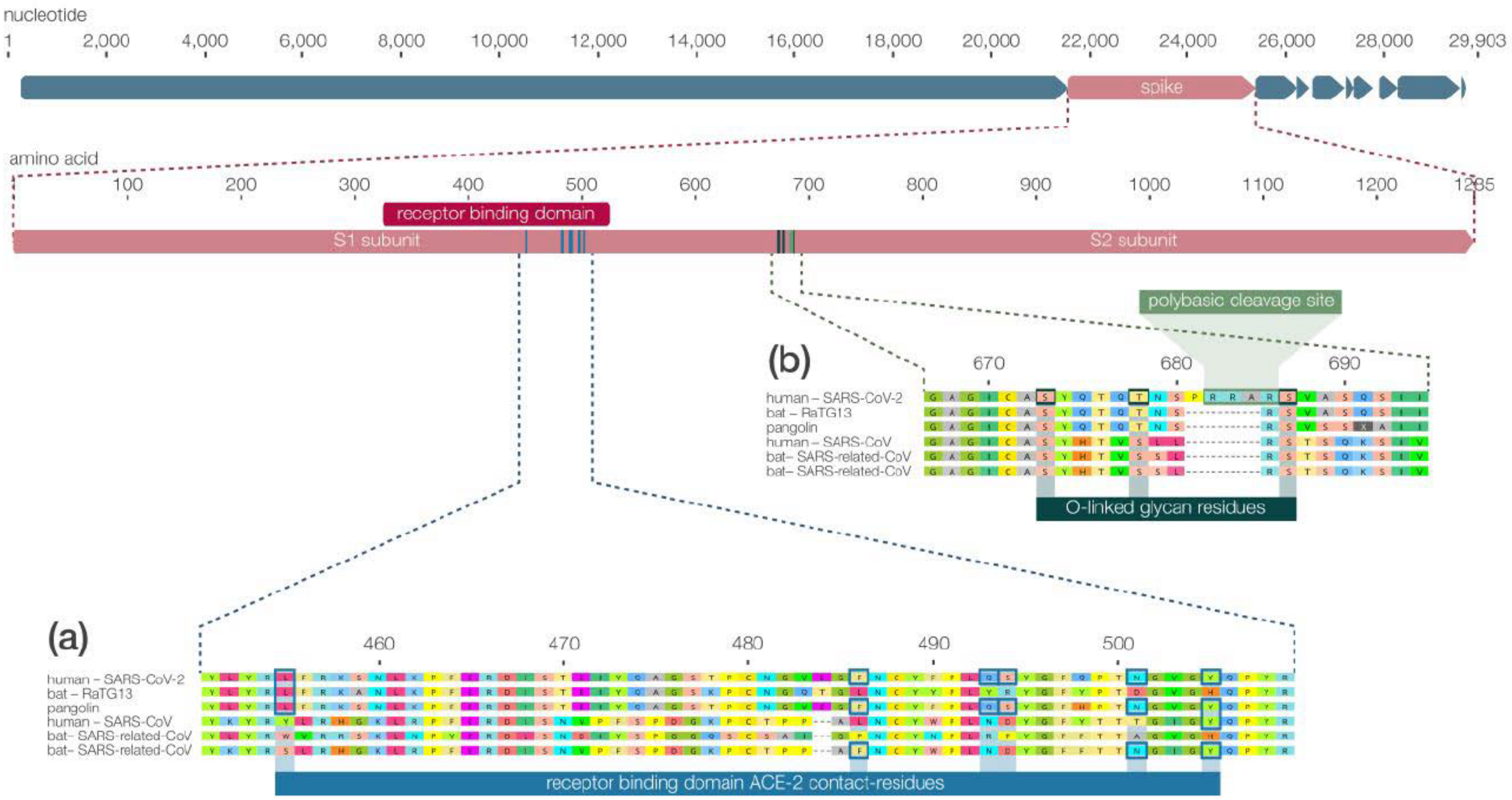

nucleotide
spike
amino acid
receptor binding domain
S1 subunit
S2 subunit
polybasic cleavage site
(b)
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
bat– SARS-related-CoV
O-linked glycan residues
(a)
human – SARS-CoV-2
bat – RaTG13
pangolin
human – SARS-CoV
bat– SARS-related-CoV
bat– SARS-related-CoV
receptor binding domain ACE-2 contact-residues

Figure Legends

Figure 1. (**a**) Mutations in contact residues of the SARS-CoV-2 spike protein. The spike protein of SARS-CoV-2 (top) was aligned against the most closely related SARS-like CoVs and SARS-CoV. Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in both SARS-CoV-2 and the SARS-CoV Urbani strain. (**b**) Acquisition of polybasic cleavage site and O-linked glycans. Both the polybasic cleavage site and the three adjacent predicted O-linked glycans are unique to SARS-CoV-2 and not previously seen in lineage B betacoronaviruses. Sequences shown are from NCBI GenBank, accession numbers MN908947, MN996532, AY278741, KY417146 and MK211376. The pangolin coronavirus sequences are a consensus generated from SRR10168377 and SRR10168378 (NCBI BioProject PRJNA573298)[29,30].

References

1. Zhou, P. *et al. Nature* (2020) doi:10.1038/s41586-020-2012-7.
2. Wu, F. *et al. Nature* (2020) doi:10.1038/s41586-020-2008-3.
3. Gorbalenya, A. E. *et al. bioRxiv* 2020.02.07.937862 (2020) doi:10.1101/2020.02.07.937862.
4. Jiang, S. *et al. Lancet* (2020) doi:10.1016/S0140-6736(20)30419-0.
5. Dong, E., Du, H. & Gardner, L. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30120-1.
6. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. *Adv. Virus Res.* **100**, 163–188 (2018).
7. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. *J. Virol.* (2020) doi:10.1128/JVI.00127-20.
8. Walls, A. C. *et al. bioRxiv* 2020.02.19.956581 (2020) doi:10.1101/2020.02.19.956581.
9. Wrapp, D. *et al. Science* (2020) doi:10.1126/science.abb2507.
10. Letko, M., Marzi, A. & Munster, V. *Nat Microbiol* (2020) doi:10.1038/s41564-020-0688-y.
11. Sheahan, T. *et al. J. Virol.* **82**, 2274–2285 (2008).
12. Nao, N. *et al. MBio* **8**, (2017).
13. Chan, C.-M. *et al. Exp. Biol. Med.* **233**, 1527–1536 (2008).
14. Follis, K. E., York, J. & Nunberg, J. H. *Virology* **350**, 358–369 (2006).
15. Menachery, V. D. *et al. J. Virol.* (2019) doi:10.1128/JVI.01774-19.
16. Alexander, D. J. & Brown, I. H. *Rev. Sci. Tech.* **28**, 19–38 (2009).
17. Ito, T. *et al. J. Virol.* **75**, 4439–4443 (2001).
18. Bagdonaite, I. & Wandall, H. H. *Glycobiology* **28**, 443–467 (2018).
19. Cui, J., Li, F. & Shi, Z.-L. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
20. Almazán, F. *et al. Virus Res.* **189**, 262–270 (2014).
21. Zhang, T., Wu, Q. & Zhang, Z. *bioRxiv* 2020.02.19.950253 (2020) doi:10.1101/2020.02.19.950253.
22. Yamada, Y. & Liu, D. X. *J. Virol.* **83**, 8744–8758 (2009).
23. http://virological.org/t/phylodynamic-analysis-90-genomes-12-feb-2020/356 (2020).
24. Huang, C. *et al. Lancet* (2020) doi:10.1016/S0140-6736(20)30183-5.
25. Dudas, G., Carvalho, L. M., Rambaut, A. & Bedford, T. *Elife* **7**, (2018).
26. Wang, N. *et al. Virol. Sin.* **33**, 104–107 (2018).

27. Ge, X.-Y. *et al. Nature* **503**, 535–538 (2013).

28. Lim, P. L. *et al. N. Engl. J. Med.* **350**, 1740–1745 (2004).

29. Wong, M. C., Javornik Cregeen, S. J., Ajami, N. J. & Petrosino, J. F. *bioRxiv* 2020.02.07.939207 (2020) doi:10.1101/2020.02.07.939207.

30. Liu, P., Chen, W. & Chen, J.-P. *Viruses* vol. 11 979 (2019).